nature genetics

Exome sequencing in sporadic autism spectrum disorders identifies severe *de novo* mutations

Brian J O'Roak[1], Pelagia Deriziotis[2], Choli Lee[1], Laura Vives[1], Jerrod J Schwartz[1], Santhosh Girirajan[1], Emre Karakoc[1], Alexandra P MacKenzie[1], Sarah B Ng[1], Carl Baker[1], Mark J Rieder[1], Deborah A Nickerson[1], Raphael Bernier[3], Simon E Fisher[2,4], Jay Shendure[1] & Evan E Eichler[1,5]

Evidence for the etiology of autism spectrum disorders (ASDs) has consistently pointed to a strong genetic component complicated by substantial locus heterogeneity[1,2]. We sequenced the exomes of 20 individuals with sporadic ASD (cases) and their parents, reasoning that these families would be enriched for *de novo* mutations of major effect. We identified 21 *de novo* mutations, 11 of which were protein altering. Protein-altering mutations were significantly enriched for changes at highly conserved residues. We identified potentially causative *de novo* events in 4 out of 20 probands, particularly among more severely affected individuals, in *FOXP1*, *GRIN2B*, *SCN1A* and *LAMC3*. In the *FOXP1* mutation carrier, we also observed a rare inherited *CNTNAP2* missense variant, and we provide functional support for a multi-hit model for disease risk[3]. Our results show that trio-based exome sequencing is a powerful approach for identifying new candidate genes for ASDs and suggest that *de novo* mutations may contribute substantially to the genetic etiology of ASDs.

ASDs are characterized by pervasive impairment in language, communication and social reciprocity and restricted interests or stereotyped behaviors[1]. Several new candidate loci for ASDs have recently been identified using genome-wide approaches that discover individually rare events of major effect[2]. A number of genetic syndromes with features of the ASD phenotype, collectively referred to as syndromic autism, have also been described[4]. Despite this progress, the genetic basis for the vast majority of cases remains unknown. Several observations support the hypothesis that the genetic basis for ASDs in sporadic cases may differ from that of families with multiple affected individuals, with the former being more likely to result from *de novo* mutation events rather than inherited variants[1,5–7]. In this study, we sequenced the protein-coding regions of the genome (the exome)[8] to test the hypothesis that *de novo* protein-altering mutations contribute substantially to the genetic basis of sporadic ASDs. In contrast with array-based analysis of large *de novo* copy number variants (CNVs), this approach has greater potential to implicate single genes in ASDs.

We selected 20 trios with an idiopathic ASD, each consistent with a sporadic ASD based on clinical evaluations (**Supplementary Table 1**), pedigree structure, familial phenotypic evaluation, family history and/or elevated parental age. Each family was initially screened by array comparative genomic hybridization (CGH) using a customized microarray[9]. We identified no large (>250 kb) *de novo* CNVs but did identify a maternally inherited deletion (~350 kb) at 15q11.2 in one family (**Supplementary Fig. 1**). This deletion has been associated with increased risk for epilepsy[10] and schizophrenia[11,12] but has not been considered causal for autism.

Similar to researchers from a previous study[13], who reported exome sequencing on ten parent-child trios with sporadic cases of moderate to severe intellectual disability, we performed exome sequencing on each of the 60 individuals separately by subjecting whole-blood derived genomic DNA to in-solution hybrid capture and Illumina sequencing (Online Methods). We obtained sufficient coverage to call variants for ~90% of the primary target (26.4 Mb) (**Table 1**). Genotype concordance with SNP microarray data was high (99.7%) (**Supplementary Table 2**), and on average, 96% of proband variant sites were also called in both parents (**Supplementary Table 3**). Given the expected rarity of true *de novo* events in the targeted exome (<1 per trio) (**Supplementary Table 4**)[14], we reasoned that most apparently *de novo* variants would result from under calling in parents or systematic false positive calls in the proband. We therefore filtered variants previously observed in the dbSNP database, 1000 Genomes Pilot Project data[15] and 1,490 other exomes sequenced at the University of Washington (**Supplementary Fig. 2**). We performed Sanger sequencing on the remaining *de novo* candidates (<5 per trio), validating 18 events within coding sequence and three additional events mapping to 3′ untranslated regions (**Table 2**). A list of predicted variant sites within these genes from the 1000 Genomes Pilot Project data[15] is provided for comparison (**Supplementary Table 5**).

[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington, USA. [2]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [3]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, Washington, USA. [4]Language and Genetics Department, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands. [5]Howard Hughes Medical Institute, Seattle, Washington, USA. Correspondence should be addressed to E.E.E. (eee@gs.washington.edu) or J.S. (shendure@uw.edu).

Received 22 February; accepted 21 April; published online 15 May 2011; doi:10.1038/ng.835

Table 1 Summary of the exome sequencing results from 20 sporadic ASD probands

Family SSC/SAGE ID[a]	Proband sex	Paternal age[b]	Maternal age[b]	Trio bases[c]	Percent target[d]	Coding SNVs[e]	Rare disruptive SNVs[f]	Coding indels[e]	Rare indels[g]	Protein coding *de novo* events
11048	M	358	358	23,901,726	88.19	14,095 (752)	131	74 (44)	27	0
11307	M	421	407	23,549,536	86.89	13,509 (583)	75	64 (40)	19	0
11580	M	443	305	23,823,712	87.90	13,912 (642)	89	62 (36)	24	1
11666	M	398	370	24,179,474	89.21	14,306 (622)	77	59 (40)	25	1
12325	M	363	313	24,088,772	88.88	13,866 (629)	79	65 (43)	24	1
12499	M	425	372	25,217,651	93.04	14,479 (634)	86	80 (47)	21	3
12575	M	351	317	24,259,870	89.51	14,568 (679)	78	80 (55)	26	0
12647	M	541	413	24,669,129	91.02	14,144 (830)	78	68 (42)	22	1
12680	M	502	471	24,437,989	90.16	14,124 (642)	69	70 (42)	24	2
12681	F	399	375	24,723,806	91.22	14,750 (691)	93	68 (39)	20	2
12817	M	485	430	24,520,475	90.47	14,364 (656)	83	72 (38)	24	2
12974	M	366	365	24,235,164	89.42	13,990 (555)	52	54 (37)	23	0
13095	M	337	322	24,460,239	90.25	14,605 (645)	66	89 (54)	29	0
13253	M	436	427	24,070,345	88.81	13,775 (610)	96	41 (25)	16	2
13284[h]	M	300	302	24,911,060	91.91	17,806 (639)	111	151 (79)	53	1
13466	M	353	385	24,676,574	91.05	14,023 (591)	72	58 (39)	23	0
13683	M	470	402	24,139,439	89.06	14,419 (725)	73	72 (49)	22	0
13708	M	397	382	23,933,169	88.30	13,997 (686)	77	78 (41)	26	2
13970	M	313	234	24,465,009	90.26	14,293 (626)	84	89 (58)	31	0
SAGE4022	F	271	283	24,130,743	89.03	14,538 (713)	141	86 (56)	29	0
Average	18M:2F	397	362	24,319,694	89.73	14,378 (658)	86	74 (45)	25	0.9

[a]Simons Simplex Collection (SSC) or Study of Autism Genetics Exploration (SAGE) family number. [b]Paternal and maternal ages (in months) at time of conception were estimated based on month-year birth information assuming a 9-month pregnancy. [c]Number of bases covered at 8× and Q30 in all three individuals. [d]±2 bp. [e]Numbers in parentheses denote variants not present in dbSNP or 1000 Genomes Project pilot data. [f]Not observed in 1,490 other exomes sequenced at the University of Washington. [g]Not equal to 3*n*. [h]Included additional RefSeq targets.

We observed subtle differences with respect to mutation rate and characteristics when compared to the previous study[13] (**Supplementary Note**). The overall protein-coding *de novo* rate (0.9 events per trio) was slightly higher than expected[14] (0.59 events per trio), suggesting that we identified the majority of the *de novo* events in these trios (**Supplementary Table 4**). The transition to transversion ratio was highly skewed (18:2), with eight transitions mapping to hypermutable CpG dinucleotides[14]. The proportion of synonymous events was higher than expected based on a neutral model and may reflect selection against embryonic lethal non-synonymous variants. We successfully determined the parent of origin for seven events, six of which occurred on the paternal haplotype (**Table 2**). Notably, the eight probands with two or more validated *de novo* events corresponded to families with higher parental age (Mann-Whitney U, combined age, one-sided $P < 0.004$).

Eleven of the 18 coding *de novo* events are predicted to alter protein function. Each of these mutations occurred in a different gene, precluding a statistical assessment for any specific locus despite their deleterious nature (for example, using PolyPhen-2)[16]. We assessed whether proband *de novo* mutations were enriched in the aggregate for disruptive events by considering two independent quantitative measures: the nature of the amino-acid replacement (Grantham matrix score[17]) and the degree of nucleotide-level evolutionary conservation (Genomic Evolutionary Rate Profiling (GERP)[18,19]) (**Fig. 1a**,**b**). For comparison, we sequenced 20 exomes from unrelated ethnically matched controls (from HapMap) and applied the same filters to identify coding-sequence mutations that were common or private to each of the samples. These control DNA samples were isolated from immortalized lymphoblasts; however, the counts of private variants in the cases and controls were highly similar, suggesting that the contribution of new somatic events is likely minimal (**Supplementary Fig. 3**).

We determined by simulation the expected mean GERP and Grantham distributions for ten randomly selected common or private control single nucleotide variants (SNVs) (Online Methods). When we compared the observed means of the ten *de novo* protein-altering ASD proband variants to the distribution of common control SNVs (**Fig. 1a**), they corresponded to more highly conserved (GERP, $P < 0.001$) and disruptive amino acid mutations (Grantham, $P = 0.015$). If we limited the analysis to the private control SNVs, which serve as a proxy for evolutionarily young mutation events (**Fig. 1b**), we again found that the *de novo* events were at the right tail of these distributions. Only the mean GERP score, however, remained significant (GERP, $P = 0.02$; Grantham, $P = 0.115$). In total, these results suggest that these *de novo* mutation sites are subjected to stronger selection and are likely to have functional impact.

We identified a subset of trios (4 out of 20) with disruptive *de novo* mutations that are potentially causative, including genes previously associated with autism, intellectual disability and epilepsy (**Table 2** and **Supplementary Note**). We examined the available clinical data for these four families and found that they were among the most severely affected individuals in our study based on intelligence quotient (IQ) measures and on calibrated severity score[20] (CSS), which is largely independent from IQ and focuses specifically on autistic features, with a score of 10 being the most severe (**Fig. 1c**,**d**). For example, in proband 12681, we identified a single-base substitution (IVS9-2A>G) at the canonical 3′ splice site of exon 10 in *GRIN2B* (encoding glutamate receptor, ionotropic, N-methyl D-aspartate 2B) (**Supplementary Fig. 4a**,**b**). This subject is severely affected (CSS 9), with evidence of early onset, possible regression and co-morbidity for mild intellectual disability. Expression and association studies have suggested that glutamatergic neurotransmission may play a role in ASDs[4]. Recently, a study[21] described *GRIN2A* and *GRIN2B* as sites of recurrent *de novo* mutations in individuals with mild to moderate intellectual disability and/or epilepsy, suggesting variable expressivity. Our data suggest that *de novo* mutations in *GRIN2B* may also lead to an ASD presentation.

Table 2 Summary of confirmed *de novo* mutation events

Proband	Type	Chromosome: position	Gene symbol	Variant	Amino acid change	GERP score	Grantham score	PolyPhen-2	CpG	Ts/Tv	Mutation origin
SNVs											
11580.p1	Missense	Chr20:2,239,665	*TGM3*	R	p.Val144Ile	5.15	29	Probably damaging	Y	Ts	Maternal
11666.p1	Missense	Chr9:132,904,111	*LAMC3*[a]	R	p.Asp339Gly	4.92	94	Probably damaging	N	Ts	Paternal
12325.p1	3′ UTR	Chr12:55,708,658	*MYO1A*	R		2.23			N	Ts	
12325.p1	Missense	Chr16:19,951,169	*GPR139*	Y	p.Ser151Gly	1.71	56	Benign	N	Ts	
12499.p1	Missense	Chr2:166,556,317	*SCN1A*[a]	R	p.Pro1894Leu	5.55	98	Probably damaging	N	Ts	Paternal
12499.p1	Synonymous	Chr3:38,033,207	*PLCD1*	K		−8.24			Y	Tv	
12499.p1	Missense	Chr6:152,865,504	*SYNE1*	Y	p.Tyr282Cys	4.48	194	Probably damaging	N	Ts	
12575.p1	3′ UTR	Chr9:32,619,906	*TAF1L*	R		−1.02			Y	Ts	
12647.p1	3′ UTR	Chr16:23,585,994	*DCTN5*	Y		−0.989			N	Ts	
12647.p1	Missense	Chr5:68,453,390	*SLC30A5*	S	p.Ser561Arg	4.6	110	Possibly damaging	N	Tv	
12680.p1	Synonymous	Chr2:101,992,478	*IL1R2*	Y		−1.53			N	Ts	
12680.p1	Synonymous	Chr5:132,251,451	*AFF4*	Y		−11.2			Y	Ts	Paternal
12681.p1	3′ splice	Chr12:13,614,220	*GRIN2B*[a]	Y		4.17	215[b]		N	Ts	Paternal
12681.p1	Synonymous	Chr7:142,274,902	*EPHB6*	Y		−3.14			Y	Ts	Paternal
12817.p1	Synonymous	Chr2:143,724,639	*ARHGAP15*	R		3.51			N	Ts	
13253.p1	Missense	Chr3:39,204,494	*XIRP1*	Y	p.Val483Met	2.04	21	Probably damaging	N	Ts	
13253.p1	Synonymous	Chr16:74,121,475	*CHST5*	Y		−3.22			Y	Ts	
13284.p1	Synonymous	Chr2:179,145,956	*TTN*	Y		0.328			Y	Ts	
13708.p1	Missense	Chr17:58,033,198	*TLK2*	Y	p.Ser595Leu	5.43	145	Probably damaging	Y	Ts	
13708.p1	Missense	Chr3:30,004,687	*RBMS3*	Y	p.Thr383Met	5.44	81	Probably damaging	Y	Ts	
Indels											
12817.p1	Frameshift	Chr3:71,132,860	*FOXP1*[a]	+T	p.Ala339Ser*fs*X4	5.38[c]	215[b]		NA	NA	Paternal

[a]Disruptive *de novo* mutations that are potentially causative. [b]Maximum Grantham score given for splice and frameshifting variants. [c]Average GERP score for two sites flanking the insertion. UTR, untranslated region.

Proband 12499 has a missense variant (p.Pro1894Leu) at a highly conserved position in *SCN1A* (encoding sodium channel, voltage-gated, type I, alpha subunit) that is predicted to be functionally deleterious (**Supplementary Fig. 4c**). This subject is severely affected (CSS 8), with evidence of early onset, possible regression, language delay, a diagnosis of epilepsy and mild intellectual disability. *SCN1A* was previously associated with epilepsy and has been suggested as an ASDs candidate gene[22,23], although limited screening has been conducted in idiopathic ASDs. Hundreds of disease-associated mutations have been described in epilepsy, and typically individuals with *de novo* events show more severe phenotypes[24]. The proband also carries the maternally inherited 15q11.2 deletion that increases the risk for epilepsy[10].

Proband 11666 has a missense variant (p.Asp399Gly) at a highly conserved position within the second laminin-type epidermal growth factor-like domain of *LAMC3* (encoding laminin, gamma 3) that is predicted to be functionally deleterious (**Supplementary Fig. 4d**). This subject is severely affected (CSS 10), with evidence of early onset and moderate intellectual disability. *LAMC3* is not known to be involved in neuronal development; however, human microarray data have shown expression in many areas of the cortex and limbic system[25]. Additional study is warranted, as laminins have structural similarities to the neurexin and contactin-associated families of proteins, both of which have been associated with ASDs[2].

The fourth example of a potentially causative mutation is a single-base insertion in *FOXP1* (encoding forkhead box P1) introducing a frameshift and premature stop codon (p.Ala339Ser*fs*X4) in proband 12817 (**Fig. 1e**). This subject is severely affected (CSS 8), with evidence for regression, language delay and co-morbidity for moderate intellectual disability and nonfebrile seizures. Recently, rare occurrences of large *de novo* deletions and a nonsense variant disrupting *FOXP1* were reported in individuals with mild to moderate intellectual disability and language defects, with or without ASD features[26,27]. *FOXP1* encodes a member of the forkhead-box family of transcription factors and is closely related to *FOXP2*, a gene implicated in rare monogenic forms of speech and language disorder[28–31]. Functional evidence of heterodimer formation and overlapping neural expression patterns suggests that FOXP1 and FOXP2 can co-regulate gene expression in the brain[32,33]. We assessed relative levels of the mutant transcript in proband-derived lymphoblasts and found strong evidence for nonsense-mediated decay (NMD) (**Supplementary Fig. 5a**). HEK293T cell-based functional assays further showed that, if translated, the protein would be truncated and mislocalized from the nucleus to the cytoplasm, similar to results obtained with *FOXP2* mutations[31] (**Supplementary Fig. 5b**,**c**).

Remarkably, in addition to the *FOXP1* mutation, proband 12817 also carried an inherited missense variant (p.His275Ala) at a highly conserved position in *CNTNAP2* (encoding contactin associated protein-like 2) that is predicted to be functionally deleterious. This variant is likely to be extremely rare or private, as it was not observed in 942 previously sequenced controls[34] or in 1,490 other exomes. *CNTNAP2* is directly downregulated by FOXP2 (ref. 35) and has been independently associated with ASD and specific language impairment[34–37]. In HEK293T cells, we found that wild-type FOXP1 significantly reduced expression of *CNTNAP2* ($P = 0.0005$), whereas the truncated protein was associated with a threefold expression increase ($P = 0.0056$) (**Supplementary Note** and **Supplementary Fig. 5d**). Overall, we hypothesize that FOXP1 haploinsufficiency (caused by NMD), combined with dysfunction of FOXP1 mutant proteins that escape this process, may result in overexpression of CNTNAP2 protein, amplifying any deleterious effects of p.His275Ala in the proband.

Among the ~110 (85 SNVs and 25 indels) previously unreported inherited protein-altering variants in each proband, we identified several rare inherited variants in genes overlapping SFARI Gene[38], a curated database of potential ASD candidate loci, but we saw no excessive

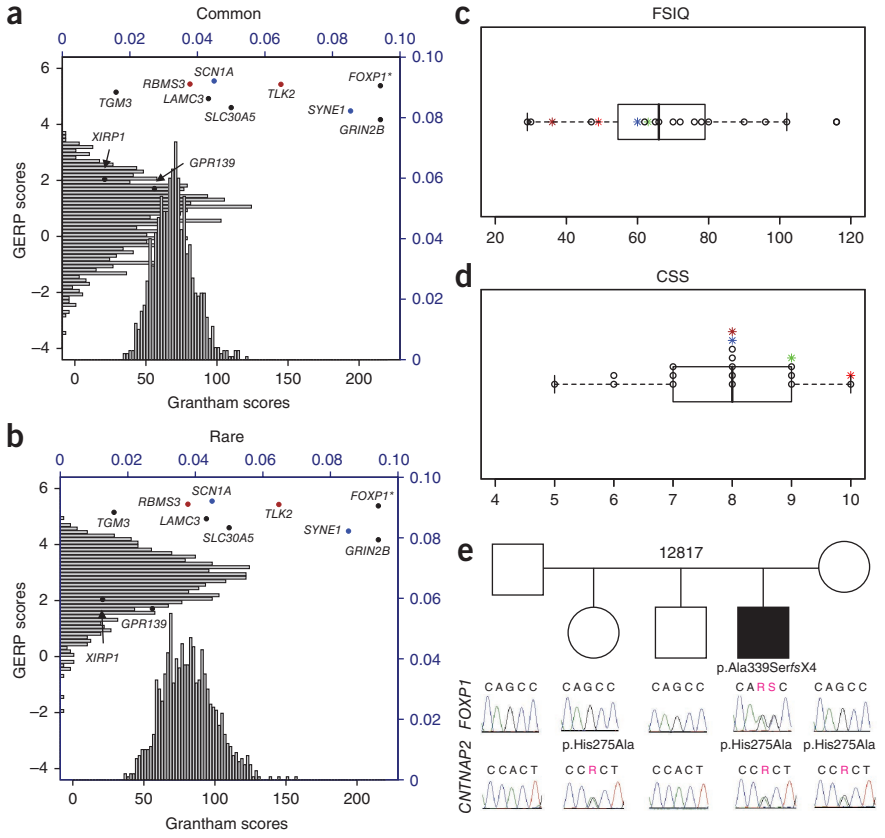

Figure 1 Evaluation of *de novo* mutations by simulation and proband severity and pedigree of family 12817. (**a**,**b**) We compared the mean Grantham (black *x* axis) and GERP scores (black *y* axis) of the ten proband *de novo* protein-changing substitutions to 20 HapMap control samples by building a distribution of the mean values of ten randomly selected common or private variants over 1,000 trials. Splice-site and nonsense events were given a maximum Grantham score (215); we did not include indels in the simulation. Histograms show the relative frequency (blue axes) of each distribution. Points show the proband variants, with variants from the same individual highlighted (blue, 13708.p1; red, 12499.p1). The proband mean values were 4.349 for GERP and 104.3 for Grantham. **FOXP1* was not included in proband mean values. (**a**) Control common variants (GERP, $P < 0.001$; Grantham, $P = 0.015$). (**b**) Control rare variants (GERP, $P = 0.02$; Grantham, $P = 0.115$). (**c**,**d**) We evaluated the disease severity of the mutation carriers 12817.p1 (*FOXP1*; brown), 12681.p1 (*GRIN2B*; green), 12499 (*SCN1A*; blue) and 11666.p1 (*LAMC3*; red). (**c**) Box and whisker plot of full scale intelligence quotient (FSIQ) values. (**d**) Box and whisker plot of calibrated severity scores (CSS) based on the autism diagnostic observation schedule (ADOS). Data were available for 19 out of 20 probands; we estimated CSS for two probands based on ADOS module 4 data. (**e**) Pedigree of family 12817 showing chromatogram traces surrounding the *FOXP1* (top) and *CNTNAP2* (bottom) mutation events. The proband carries a *de novo* single-base (+A relative to mRNA) frameshifting mutation (resulting in p.Ala339Ser*fs*X4) in *FOXP1* and an inherited missense variant (p.His275Ala) in *CNTNAP2*.

burden in cases relative to controls (**Supplementary Table 6**). Although the numbers of subjects from our pilot study are few, we did observe two cases with a notable *de novo* event and a potential inherited risk variant (12817.p1, *FOXP1* and *CNTNAP2*; 12499.p1, *SCN1A* and 15q11.2 deletion), highlighting that in some sporadic families a multi-hit model may be playing a role[3] (**Supplementary Table 7**). In the future, this hypothesis could be further explored by comparing burden in a much larger number of affected-unaffected sibling pairs.

The probands with the four potentially causative *de novo* events met strict criteria for a diagnosis of autistic disorder (**Supplementary Note**). Our finding of *de novo* events in genes that have also been disrupted in children with intellectual disability without ASD, intellectual disability with ASD features or epilepsy provides further evidence that these genetic pathways may lead to a spectrum of neurodevelopmental outcomes depending on the genetic and environmental context[2,4]. Recent data suggest that CNVs may also blur these lines, with diverse conditions all showing association to the same loci[2,4]. Distinguishing primary from secondary effects will require a better understanding of the underlying biology and identification of interacting genetic and environmental factors within the phenotypic context of the family. The identification of *de novo* events along with disruptive inherited mutations underlying 'sporadic' ASDs has the potential to fundamentally transform our understanding of the genetic basis of autism.

URLs. SeattleSeq server, http://gvs.gs.washington.edu/SeattleSeqAnnotation/; Primer3, http://frodo.wi.mit.edu/primer3/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Accession numbers. Reference sequences are available from the NCBI CCDS database under the following accession codes: *CNTNAP2*, CCDS5889.1; *FOXP1*, CCDS2914.1; *GRIN2B*, CCDS8662.1; *LAMC3*, CCDS6938.1; *SCN1A*, CCDS33316.1.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We would like to thank and recognize the following ongoing studies that produced and provided exome variant calls for comparison: National Heart, Lung, and Blood Institute (NHBLI) Lung Cohort Sequencing Project (HL 1029230), NHLBI Women's Health Initiative (WHI) Sequencing Project (HL 102924), National Institute of Environmental Health Sciences (NIEHS) SNPs (HHSN273200800010C), NHLBI/National Human Genome Research Institute (NHGRI) SeattleSeq (HL 094976), NHGRI Next Generation Mendelian Genetics (HG 005608) and the Northwest Genomics Center (HL 102926). We also thank M.-C. King and S. Stray for processing and managing DNA samples, B.H. King and E. Bliss for their work in subject recruitment and phenotype collection, E. Turner, C. Igartua, I. Stanaway, M. Dennis and B. Coe for thoughtful discussions, M. State for providing SNP genotyping data and especially the families that volunteered their time to participate in this research. This work was supported by US National Institutes of Health grant HD065285 (E.E.E. and J.S.), Wellcome Trust core award 075491/Z/04 (S.E.F. and P.D.), the Max Planck Society (S.E.F.) and grants from the Simons Foundation Autism Research Initiative (SFARI) (191889, 137578 and 137593) (E.E.E., R.B., S.E.F. and P.D.). E.E.E. is an Investigator of the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

E.E.E., J.S. and B.J.O. designed the study and drafted the manuscript. E.E.E. and J.S. supervised the study. R.B. analyzed the clinical information and contributed

to the manuscript. S.E.F. and P.D. designed cell-based functional experiments, analyzed data, interpreted results and contributed to the manuscript. S.G., C.B. and L.V. generated and analyzed array CGH data. C.L. performed Illumina GAIIx sequencing. B.J.O. and E.K. developed the analysis pipeline and analyzed sequence data. A.P.M. and S.B.N. designed and optimized capture protocol. B.J.O., L.V., A.P.M. and S.B.N. constructed exome libraries. B.J.O., L.V., A.P.M. and J.J.S. performed mutation validation and haplotype characterization. B.J.O. and J.J.S. performed the evaluation of 12817 lymphoblast cell lines. P.D. performed functional experiments. M.J.R. and D.A.N. performed sequencing of control samples.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturegenetics/.

Published online at http://www.nature.com/naturegenetics/.

1. Bailey, A. *et al.* Autism as a strongly genetic disorder: evidence from a British twin study. *Psychol. Med.* **25**, 63–77 (1995).
2. O'Roak, B.J. & State, M.W. Autism genetics: strategies, challenges, and opportunities. *Autism Res.* **1**, 4–17 (2008).
3. Girirajan, S. *et al.* A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nat. Genet.* **42**, 203–209 (2010).
4. Abrahams, B.S. & Geschwind, D.H. Advances in autism genetics: on the threshold of a new neurobiology. *Nat. Rev. Genet.* **9**, 341–355 (2008).
5. Sebat, J. *et al.* Strong association of *de novo* copy number mutations with autism. *Science* **316**, 445–449 (2007).
6. Marshall, C.R. *et al.* Structural variation of chromosomes in autism spectrum disorder. *Am. J. Hum. Genet.* **82**, 477–488 (2008).
7. Durkin, M.S. *et al.* Advanced parental age and the risk of autism spectrum disorder. *Am. J. Epidemiol.* **168**, 1268–1276 (2008).
8. Ng, S.B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
9. Bailey, J.A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
10. de Kovel, C.G. *et al.* Recurrent microdeletions at 15q11.2 and 16p13.11 predispose to idiopathic generalized epilepsies. *Brain* **133**, 23–32 (2010).
11. Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455**, 232–236 (2008).
12. Kirov, G. *et al.* Support for the involvement of large CNVs in the pathogenesis of schizophrenia. *Hum. Mol. Genet.* **18**, 1497–1503 (2009).
13. Vissers, L.E. *et al.* A *de novo* paradigm for mental retardation. *Nat. Genet.* **42**, 1109–1112 (2010).
14. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci. USA* **107**, 961–968 (2010).
15. Durbin, R.M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
16. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
17. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185**, 862–864 (1974).
18. Cooper, G.M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).
19. Cooper, G.M. *et al.* Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat. Methods* **7**, 250–251 (2010).
20. Gotham, K., Pickles, A. & Lord, C. Standardizing ADOS scores for a measure of severity in autism spectrum disorders. *J. Autism Dev. Disord.* **39**, 693–705 (2009).
21. Endele, S. *et al.* Mutations in *GRIN2A* and *GRIN2B* encoding regulatory subunits of NMDA receptors cause variable neurodevelopmental phenotypes. *Nat. Genet.* **42**, 1021–1026 (2010).
22. Claes, L. *et al. De novo* mutations in the sodium-channel gene *SCN1A* cause severe myoclonic epilepsy of infancy. *Am. J. Hum. Genet.* **68**, 1327–1332 (2001).
23. Weiss, L.A. *et al.* Sodium channels *SCN1A*, *SCN2A* and *SCN3A* in familial autism. *Mol. Psychiatry* **8**, 186–194 (2003).
24. Mulley, J.C. *et al. SCN1A* mutations and epilepsy. *Hum. Mutat.* **25**, 535–542 (2005).
25. Lein, E.S. *et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168–176 (2007).
26. Hamdan, F.F. *et al. De novo* mutations in *FOXP1* in cases with intellectual disability, autism, and language impairment. *Am. J. Hum. Genet.* **87**, 671–678 (2010).
27. Horn, D. *et al.* Identification of *FOXP1* deletions in three unrelated patients with mental retardation and significant speech and language deficits. *Hum. Mutat.* **31**, E1851–E1860 (2010).
28. Lai, C.S., Fisher, S.E., Hurst, J.A., Vargha-Khadem, F. & Monaco, A.P. A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature* **413**, 519–523 (2001).
29. Feuk, L. *et al.* Absence of a paternally inherited *FOXP2* gene in developmental verbal dyspraxia. *Am. J. Hum. Genet.* **79**, 965–972 (2006).
30. MacDermot, K.D. *et al.* Identification of *FOXP2* truncation as a novel cause of developmental speech and language deficits. *Am. J. Hum. Genet.* **76**, 1074–1080 (2005).
31. Vernes, S.C. *et al.* Functional genetic analysis of mutations implicated in a human speech and language disorder. *Hum. Mol. Genet.* **15**, 3154–3167 (2006).
32. Li, S., Weidenfeld, J. & Morrisey, E.E. Transcriptional and DNA binding activity of the Foxp1/2/4 family is modulated by heterotypic and homotypic protein interactions. *Mol. Cell. Biol.* **24**, 809–822 (2004).
33. Teramitsu, I., Kudo, L.C., London, S.E., Geschwind, D.H. & White, S.A. Parallel FoxP1 and FoxP2 expression in songbird and human brain predicts functional interaction. *J. Neurosci.* **24**, 3152–3163 (2004).
34. Bakkaloglu, B. *et al.* Molecular cytogenetic analysis and resequencing of contactin associated protein-like 2 in autism spectrum disorders. *Am. J. Hum. Genet.* **82**, 165–173 (2008).
35. Vernes, S.C. *et al.* A functional genetic link between distinct developmental language disorders. *N. Engl. J. Med.* **359**, 2337–2345 (2008).
36. Arking, D.E. *et al.* A common genetic variant in the neurexin superfamily member *CNTNAP2* increases familial risk of autism. *Am. J. Hum. Genet.* **82**, 160–164 (2008).
37. Alarcón, M. *et al.* Linkage, association, and gene-expression analyses identify *CNTNAP2* as an autism-susceptibility gene. *Am. J. Hum. Genet.* **82**, 150–159 (2008).
38. Banerjee-Basu, S. & Packer, A. SFARI Gene: an evolving database for the autism research community. *Dis. Model Mech.* **3**, 133–135 (2010).

ONLINE METHODS

ASD samples. We examined 20 families where there was a single child with a diagnosis of ASD obtained from the Simons Simplex Collection (SSC)[39] ($n = 19$) and the Study of Autism Genetics Exploration (SAGE) ($n = 1$). We required subjects to have at least one unaffected sibling with no history of ASD or intellectual disability based on medical history and/or pedigree evaluation. Parents showed no signs of a broader autism phenotype based on Broad Autism Phenotype Questionnaire[40] (cutoff ≤ 3) and Adult-Social Responsiveness Scale (SRS)[41] (cutoff ≤ 40). Siblings were discordant for the SRS and Vineland-II, with the unaffected sibling falling within the normal range.

Control exomes. Exome sequence data from 20 unrelated CEPH Utah samples of European ancestry were obtained from immortalized lymphoblast extracted genomic DNA and processed through the same variant calling pipeline.

Control simulation. Protein-altering control variants were randomly shuffled, and ten were selected over 1,000 trials to build an expected distribution of common or rare events, computing the mean Grantham and GERP scores for each trial. Empirical *P* values were then calculated based on the number of trials in which the control mean was greater than the observed mean of the ten proband disruptive *de novo* variants.

Other control exomes. Variants were screened against exome calls from 1,490 samples to remove common variants and systematic artifacts. Samples were obtained from the NHLBI Lung Cohort Sequencing Project ($n = 635$), the NHLBI WHI Sequencing Project ($n = 744$), NIEHS SNPs ($n = 87$) and NHLBI/NHGRI SeattleSeq and NHGRI Next Generation Mendelian Genetics ($n = 24$).

Array CGH. We designed custom targeted arrays comprising 135,000 probes (Roche NimbleGen) with higher density probe coverage (mean probe spacing 2.5 kb) in genomic hotspots and a lower probe density in genomic backbone (with mean probe spacing of 38 kb). Microarray hybridization experiments and CNV analysis were performed as described previously[42,43] using a single unaffected male reference (GM15724 from Coriell).

ASD exome library construction. Exome shotgun libraries were prepared as described previously[44] from 3 µg whole-blood extracted genomic DNA with modifications for solution-based capture using either a standard or a barcoding protocol. Each member of a trio received matching protocols. In the standard protocol, pre-capture PCR was performed in four reactions per sample, with each reaction in a 40 µl volume with 10 µl of library, 1× iProof High Fidelity Master Mix (Bio-Rad) and 0.625 µM of each primer under the following conditions: 96 °C for 2 min, 16 cycles of 96 °C for 20 s, 63 °C for 30 s and 72 °C for 45 s and, finally, 72 °C for 5 min. Reactions for each sample were pooled and column purified (PCR Purification Kit, Qiagen). In the barcoded protocol, pre-capture PCR was performed in five reactions per sample, with each reaction in a 40 µl volume with 10 µl of library, 1× iProof High Fidelity Master Mix (Bio-Rad) and 0.625 µM of each primer under the following conditions: 98 °C for 30 s, six cycles of 98 °C for 10 s, 60 °C for 30 s and 72 °C for 45 s and, finally, 72 °C for 5 min. Reactions for each sample were pooled and bead purified (AMPure, Agencourt).

Library capture and sequencing. For each library, 1 µg was hybridized for 72 h at 47 °C with SeqCap EZ Exome probes v1.0 (Roche) as per manufacturer's protocols and blocked with 100 µl human Cot1 DNA (Invitrogen) and 10 µl each blocker primer at 100 µM (**Supplementary Table 8**). For standard preps pre-sequencing PCR, ten reactions were performed per sample, with each reaction in a 50 µl volume with 4 µl of library, 1× iProof High Fidelity Master Mix (Bio-Rad) and 0.625 µM of each primer under the following conditions: 98 °C for 30 s, 20 cycles of 98 °C for 10 s, 60 °C for 30 s and 72 °C for 30 s and, finally, 72 °C for 5 min. Reactions for each sample were pooled and column purified. For barcode preps, five reactions were performed per sample, with each reaction in a 50 µl volume with 10 µl of library as above, but for only 13 cycles. Reactions for each sample were pooled and bead purified. Family 13284 was also captured using a custom RefSeq solid state array (Agilent) as described[45]. Libraries were sequenced on an Illumina Genome Analyzer IIx according to the manufacturer's instructions for single or paired-end 76-bp reads.

Estimating the expected mutation rate. We used mutation rates published by Lynch[14], as he provides an estimate of substitution, insertion and deletion rates and is in close agreement with other recent estimates[15,46]. We first calculated the number of callable bases in the first six completed trios that directly overlapped or were within 2 bp of the ends of the exon coordinates from the current hg18 build of the Consensus Coding Sequence (CCDS) project. This analysis yielded 22,546,796 coding and 523,843 splice bases covered in an average trio. Using the following haploid mutation rates (substitution rate of 1.28×10^{-8}, deletion rate of 5.80×10^{-10} and insertion rate of 2.00×10^{-10}), we calculated total number of substitution, indel and splice-site events. Finally, we separated the expected substitutions based their probably of causing an amino acid change under a random model of 25% synonymous, 71.25% missense and 3.75% nonsense substitutions.

Exome data analysis for SNVs and small indels. The exome definition was based on consensus coding sequence (CCDS 2009) of the human reference genome (build36). For *BWA* (0.5.6)[47] mapped reads, we removed reads with mapping scores of zero, incorrect or unmapped pairs and potential duplicates. Consensus genotypes were generated using SAMtools (0.1.7)[48]. Variant positions were then pulled and filtered using the samtools.pl varFilter (all defaults except, -D 10000 and awk '($3 = = "*"&&$6> = 50) || ($3! = "*"&&$6> = 20)'). Variants were then run through a custom pipeline, Haystack, to identify Mendelian errors. For SNVs, proband variants were filtered to 8× and consensus or SNP quality 30 positions that were called in the entire trio. Possibly *de novo* events were then compared against 1,490 other exomes sequenced at University of Washington to remove systematic artifacts and rare population variants. Variants were then annotated using the SeattleSeq server (see URLs). Possibly *de novo* events were manually inspected using the IGV browser[15]. Base pair positions where >10% of the parental reads were concordant with the putative *de novo* event were excluded. We found that approximately one-third of on-target *de novo* calls mapped to exons within segmental duplications[49] and likely represent paralogous sequence variants from missing duplicated genes or variants of limited functional consequence[50]. For *BWA* called indels, variants were filtered to a high confidence set supported by at least 30% of the reads with a minimum of eight variant reads. Positions without an indel call or 8× and consensus or SNP quality 30 position in both parents were removed. Variants were then annotated using a custom script, compared to dbSNP129 and dbSNP130 and 1000 Genomes Pilot indels and manually inspected.

Exome data analysis for large indels and CNVs. We used a general split-read–based combinatorial algorithm (E.K., C. Alkan, B.J.O., D.A.N. & E.E.E., unpublished data) to detect structural variation in the form of insertions, deletions and copy number variants ranging in length from 1 bp to 1 Mb. This method identifies one-end anchored placements using *mrsFAST*[51] alignments. The unmapped ends of the one-end anchored reads are split into pieces at the breakpoints of possible indels. Our method computes the exact breakpoints of these splits and aims to map them back to the reference exome efficiently. The final set of indels is reported based on these consistent split mappings around these events within basepair resolution.

***De novo* validation.** Possibly *de novo* variants were tested by designing custom primers with Primer3 (see URLs), performing standard PCR reactions using 10 ng of DNA from the proband, father, mother and for 19 out of 20 families an unaffected sibling, followed by Sanger sequencing.

Family 12817 complementary DNA (cDNA) analysis. Transformed live lymphoblast cell lines were obtained on the father, mother, proband and two siblings from family 12817. Cultures were grown as described[52] and split into two flasks, one of which was supplemented with 100 µg/ml of emetine (Sigma) to inhibit NMD[52]. After 7 h of incubation, total RNA was extracted using RNeasy Mini Kit (Qiagen). First strand cDNA was generated using Transcriptor High Fidelity cDNA Synthesis Kit (Roche). *FOXP1* cDNA was amplified using standard PCR with primers spanning several exons and subjected to Sanger sequencing.

Functional characterization of FOXP1. Plasmids expressing full-length FOXP2 and FOXP1, as well as FOXP2.R553H and the FOXP2.10+ were

doi:10.1038/ng.835

described elsewhere[31]. The FOXP1mut construct was directly synthesized by GenScript USA Inc. All constructs were expressed from the pcDNA4/HisMax vector and were in frame with an N-terminal vector-encoded Xpress tag. The anti-Xpress mouse monoclonal antibody (Invitrogen) was used to detect protein expression of all FOXP constructs. β-actin mouse monoclonal antibody (Sigma) was used as loading control. HEK293T cells were cultured as described[31] and transfected with either empty pcDNA4/HisMax vector or with pcDNA4/HisMax FOXP constructs using Genejuice (Novagen) transfection reagent according to the manufacturer's instructions. Forty-eight hours after transfection, cells were either lysed and resolved by SDS-PAGE or seeded on coverslips for immunofluorescence, as previously described[31]. For quantitative PCR, total RNA was isolated using an RNeasy Mini Kit (Qiagen) following the manufacturer's instructions 48 h post transfection. Five hundred nanograms total RNA was reverse-transcribed using the Superscript III Reverse Transcriptase and random primers (Invitrogen). Quantitative real-time PCR was performed on an iQ5 (Bio-Rad) (**Supplementary Table 9**).

39. Fischbach, G.D. & Lord, C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68**, 192–195 (2010).
40. Hurley, R.S., Losh, M., Parlier, M., Reznick, J.S. & Piven, J. The broad autism phenotype questionnaire. *J. Autism Dev. Disord.* **37**, 1679–1690 (2007).
41. Constantino, J.N. & Todd, R.D. Intergenerational transmission of subthreshold autistic traits in the general population. *Biol. Psychiatry* **57**, 655–660 (2005).
42. Selzer, R.R. *et al.* Analysis of chromosome breakpoints in neuroblastoma at sub-kilobase resolution using fine-tiling oligonucleotide array CGH. *Genes Chromosom. Cancer* **44**, 305–319 (2005).
43. Itsara, A. *et al.* Population analysis of large copy number variants and hotspots of human genetic disease. *Am. J. Hum. Genet.* **84**, 148–161 (2009).
44. Igartua, C. *et al.* Targeted enrichment of specific regions in the human genome by array hybridization. *Curr. Protoc. Hum. Genet.* **Chapter 18**, Unit 18 3 (2010).
45. Ng, S.B. *et al.* Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome. *Nat. Genet.* **42**, 790–793 (2010).
46. Roach, J.C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).
47. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
48. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
49. Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J. & Eichler, E.E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.* **11**, 1005–1017 (2001).
50. Sudmant, P.H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010).
51. Hach, F. *et al.* mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat. Methods* **7**, 576–577 (2010).
52. Andrés, A.M. *et al.* Balancing selection maintains a form of *ERAP2* that undergoes nonsense-mediated decay and affects antigen presentation. *PLoS Genet.* **6**, e1001157 (2010).